Report of the Baseline Model: Wikipedia Search Engine using TF-IDF and Jaccard Coefficient

Introduction: This baseline model aims to build a search engine to retrieve the most relevant Wikipedia articles based on a search query. The search engine uses two techniques: TF-IDF and Jaccard Coefficient.

Methodology: The baseline model is implemented in Python. The following steps are involved in the implementation of the project:

1. Data Collection: The data is collected using the Wikipedia API. The search query and language are given as input to the API, and it returns the search results as JSON data.
2. Data Preprocessing: The text data obtained from the API is preprocessed by removing the HTML tags, punctuations, stop words, and numbers. The text data is then tokenized into words.
3. TF-IDF Matrix Generation: The TF-IDF matrix is generated for the preprocessed text data. The TF-IDF matrix is a numerical representation of the text data that represents the importance of each word in the documents. The matrix is generated using the TfidfVectorizer class from the scikit-learn library.
4. Search using the Jaccard Coefficient: The search query is tokenized into words, and a Jaccard coefficient is calculated between the query tokens and the tokens of each document. The Jaccard coefficient measures the similarity between two sets of words. The documents with the highest Jaccard coefficients are considered the most relevant and are returned as search results.
5. Results Display: The search results are displayed in the console. The top k results are displayed, where k is a user-defined parameter.

Results: The search engine was tested with various search queries, and the search results were compared to those obtained from the Wikipedia API. The search engine was found to retrieve the most relevant Wikipedia articles for the given search query. The Jaccard coefficient was found to be an effective measure of similarity for this task.

Conclusion: While the Wikipedia search engine using TF-IDF and Jaccard coefficient effectively retrieves the most relevant Wikipedia articles for a given search query, it has certain limitations compared to the Wikipedia search engine using BERT.

1. Accuracy: The BERT-based search engine is more accurate in retrieving the most relevant Wikipedia articles as it uses a deep learning model that has been trained on a large corpus of data. The TF-IDF and Jaccard coefficient approach is a more straightforward technique that does not consider the text's semantics.

2. Speed: The BERT-based search engine may take longer to generate results due to the complexity of the model. The TF-IDF and Jaccard coefficient approach is faster as it involves simple matrix operations.
3. User Feedback: The BERT-based search engine can incorporate user feedback to improve the search results over time. The TF-IDF and Jaccard coefficient approach does not have this capability.
4. Multilingual Support: The BERT-based search engine can handle multiple languages, whereas the TF-IDF and Jaccard coefficient approach is limited to one language.

While the TF-IDF and Jaccard coefficient approach effectively retrieves the most relevant Wikipedia articles, it may not be as accurate or versatile as the BERT-based search engine.